[18] Hayashi, M., and S. Spiegelman, these PROCEEDINGS, **47**, 1564 (1961).
[19] Grossman, L., S. S. Levine, and W. S. Allison, *J. Mol. Biol.*, **3**, 47 (1961).
[20] Mandell, J. D., and A. D. Hershey, *Anal. Biochem.*, **1**, 66 (1960).
[21] Hall, C. E., E. C. Maclean, and I. Tessman, *J. Mol. Biol.*, **1**, 192 (1959).
[22] Guthrie, G. D., and R. L. Sinsheimer, *J. Mol. Biol.*, **2**, 297 (1960).
[23] Hayashi, M., and S. Spiegelman, in manuscript (1963).
[24] Kano-Sueoka, T., and S. Spiegelman, these PROCEEDINGS, **48**, 1942 (1962).
[25] Nygaard, A. P., and B. D. Hall, preprint (1963).
[26] Hayashi, M., M. N. Hayashi, and S. Spiegelman, in manuscript (1963).
[27] Strezloff, E., and F. J. Ryan, *Biochem. Biophys. Res. Comm.*, **7**, 471 (1962).
[28] Champe, S. P., and S. Benzer, these PROCEEDINGS, **48**, 532 (1962).

PRIMARY STRUCTURE AND EVOLUTION OF CYTOCHROME C

BY E. MARGOLIASH

BIOCHEMICAL RESEARCH DEPARTMENT, ABBOTT LABORATORIES, NORTH CHICAGO

Communicated by C. B. Anfinsen, August 29, 1963

Following the determination of the amino acid sequence of horse heart cytochrome *c*,[1–7] the primary structures of six other cytochromes *c* have been reported, namely, those of man,[8] pig,[9] rabbit,[9] chicken,[9] tuna,[10] and baker's yeast.[11] This work provides the most extensive series of complete amino acid sequences so far available for a set of homologous proteins from different species. The present paper considers the information yielded by a comparison of these structures, with regard to the evolution of cytochrome *c*. Some of the features of this protein from horse, man, pig, and chicken have recently been discussed,[12] particularly in relation to structural aspects of cytochrome *c* function.

Similarities.—The outstanding feature of the amino acid sequences of the cytochromes *c* considered here is the extensive degree of similarity among them. As shown in Figure 1, as much as 53% of all residues are identical in all seven proteins. When yeast cytochrome *c* is excluded from the comparison, this value increases to 73%. The six vertebrate proteins consist of a single peptide chain 104 residues long, having an acetylated amino-terminal glycyl residue and the heme attached by thioether bonds to cysteines in positions 14 and 17. Only the yeast protein has a nonacylated amino-terminal residue, carrying in place of the acetyl group a sequence of five amino acids preceding the amino-terminal glycyl residue of the vertebrate cytochromes *c*.[11] Yeast cytochrome *c* also lacks an amino acid corresponding to position 102 or 103 and thus has an over-all chain length of 108 residues.[11] Preliminary evidence indicates that the cytochrome *c* from an invertebrate, that of the moth *Samia cynthia*, also falls within the same set of homologous proteins. Like the yeast protein it lacks an acetyl group at the amino-terminal residue, bearing instead several amino acids, and has a peptide chain longer than 104 residues.[13]

The similarities of sequence in the vicinity of the heme attachments, originally noted by Tuppy and collaborators (see Paléus and Tuppy[14]), are present in all these cytochromes *c*. They consist of a pair of cysteines (nos. 14 and 17) spaced by two residues, with a basic amino acid, either lysine, as in most cases, or arginine, as in the yeast,[14] silkworm,[14] and moth [13] proteins, preceding the first cysteine, and a

Acetyl.**Gly**.*Asp*.Val.Glu.**Lys**.**Gly**.*Lys*.*Lys*.Ileu.**Phe**.Val.GluNH$_2$.*Lys*.**CyS**.Ala.*GluNH*$_2$.**CyS**.**His**.**Thr**.
10 └—HEME—┘
Val.**Glu**.Lys.**Gly**.**Gly**.*Lys*.**His**.**Lys**.Thr.**Gly**.**Pro**.**Asp**NH$_2$.**Leu**.His.**Gly**.*Leu*.**Phe**.**Gly**.**Arg**.*Lys*.*Thr*.
20 30 40
Gly.**GluNH$_2$**.**Ala**.Pro.**Gly**.Phe.Thr.**Tyr**.**Thr**.Asp.**Ala**.**Asp**NH$_2$.*Lys*.AspNH$_2$.**Lys**.*Gly*.*Ileu*.Thr.**Try**.Lys.
50 60
Glu.Glu.*Thr*.*Leu*.*Met*.Glu.Tyr.**Leu**.*Glu*.**Asp**NH$_2$.**Pro**.**Lys**.**Lys**.**Tyr**.**Ileu**.**Pro**.**Gly**.**Thr**.**Lys**.**Met**.*Ileu*.
70 80
Phe.Ala.**Gly**.*Ileu*.**Lys**.**Lys**.*Lys*.Thr.*Glu*.**Arg**.Glu.**Asp**.**Leu**.Ileu.*Ala*.**Tyr**.**Leu**.**Lys**.**Lys**.Ala.Thr.
90 100
AspNH$_2$.GluCOOH
104

FIG. 1.—Amino acid sequence of horse heart cytochrome *c*. The residues in **bold-face** type are those that are identical in the cytochromes *c* from horse, man, pig, rabbit, chicken, tuna, and baker's yeast. The residues in *italics* are those that are identical in the vertebrate proteins only.

His.Thr sequence following the second cysteine. Identical sequences are, however, found in similar situations in proteins as different from the "mammalian type" of cytochrome *c*[15] as cytochrome c_2 from *Rhodospirillum rubrum*[14] and the cytochromoid of Chromatium.[16] These regularities near the site of heme binding must therefore represent structural requirements for the attachment of the prosthetic group and the provision of at least part of its immediate environment, regardless of the functional activities of the protein.

Two areas of the amino acid sequence remain invariant. The first, extending from residues 17 to 21, immediately follows the heme attachment; it contains one of the histidines most probably involved in hemochrome formation and may well be part of the "crevice"[17] structure which enfolds the prosthetic group. No function can as yet be ascribed to the other invariant area found between residues 70 and 80. Numerous other residues and shorter sequences remain identical throughout (see Fig. 1).

The most plausible interpretation of the abundant structural similarities is that all these proteins derive phylogenetically from a common primordial cytochrome *c*, even though the species examined are as divergent as yeast and the higher mammals. Alternatively, one might consider that the observed identities result from a process of convergent evolution directed by selective pressures for structural features necessary for cytochrome *c* function. However, such an explanation is improbable. Indeed, active sites of proteins comprise only a few residues. The remainder of the peptide chain is necessary to form a particular spatial structure, a requirement known to be compatible with a variety of amino acid sequences, as must be the case, for example, with the seal[18] and sperm whale[19] myoglobins. There is thus little doubt that the seven cytochromes *c* under discussion are truly *homologous* structures in the evolutionary sense.

If, as seems likely, this conclusion can be extended to a large variety of vertebrate and invertebrate species, as well as to the plant kingdom, strong support would be obtained for the speculation[20] that living matter was effectively formed only once, within the confines of our planet, all living forms deriving from a common precursor.

In addition to maintaining the general similarities of structure, evolution appears to have exerted a very considerable, if not absolute, degree of conservatism with respect to basic and hydrophobic residues. As originally noted in the amino acid sequence of horse heart cytochrome *c*, the basic residues, lysine, arginine, and histidine on the one hand, and the hydrophobic residues, tyrosine, phenylalanine,

tryptophan, leucine, isoleucine, valine, and methionine on the other, tend to form clusters along the peptide chain.[6] Figure 2 presents the composite distribution of both these types of residues in the seven cytochromes *c* and demonstrates that these groupings have to a large extent been maintained. There are seven basic and eight hydrophobic clusters, which among the vertebrate proteins comprise all except 3–5 basic and 2 hydrophobic residues, out of totals of 22–24 and 26–29 such amino acids, respectively. Although the yeast protein differs considerably from the vertebrate cytochromes *c*, similar groupings at about the same positions along the peptide chain also occur in it. In most of the relatively few cases in which hydrophobic have replaced nonhydrophobic residues, the exchanges merely serve to complete or extend the hydrophobic groupings (see Fig. 2). It is at present impossible to decide

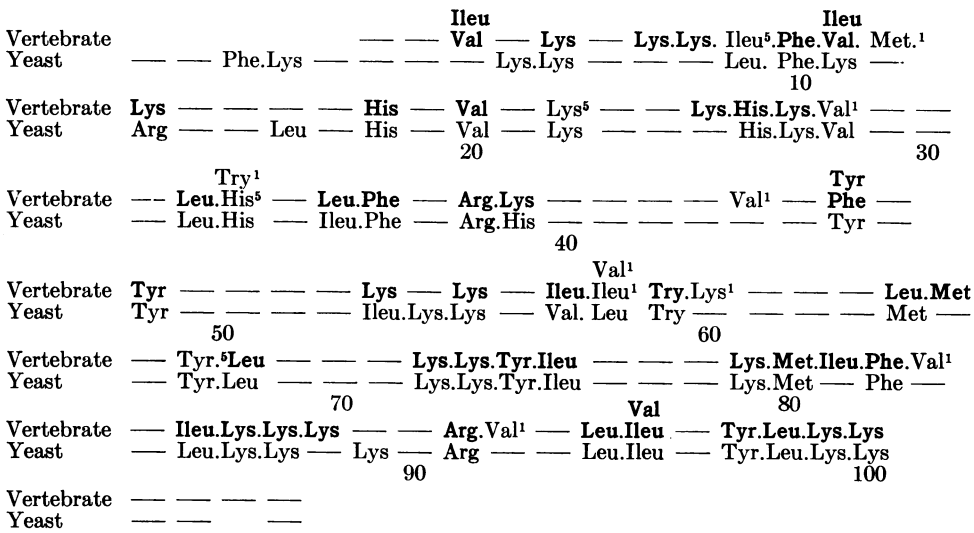

FIG. 2.—Composite distribution of basic and hydrophobic residues in the cytochromes *c* from horse, man, pig, rabbit, chicken, tuna, and baker's yeast. In positions in which only basic or only hydrophobic amino acids occur, all residues are in **bold-face** type. In other cases, the superscripts indicate the number of proteins in which the particular residues are present.

whether this conservatism is based solely on evolutionary selection for structural features of functional importance, or whether conservatism is built into the mechanism of genetic variability, mutations involving basic or hydrophobic residues tending to yield the same types of amino acids far more often than not. However, without exception, all the ribonucleic acid nucleotide triplets so far observed which can "code" for hydrophobic amino acids contain uridylic acid, while at least one of the triplet "codes" for each of the other residues does not involve this nucleotide.[21] If this finding is confirmed, it would tend to suggest that some types of evolutionary conservatism may indeed be based on the mechanism of mutation, rather than on evolutionary selection.

Similar considerations must be taken into account in attempts to ascribe functional importance to certain areas of the primary structure of a set of homologous proteins solely on the basis of their invariance over a large range of the evolutionary scale. Thus, for example, the available information does not make it possible to

decide whether a section, such as that extending from residue 70 to residue 80, has remained invariant as a result of strict functional requirements or whether such constancy merely reflects a particular stability of the genetic material corresponding to this sequence. Indeed, the presence of apparent genetic "hot spots" (see below) implies genetic "cold spots." Invariant residues and sequences could be an expression of properties of the corresponding deoxyribonucleic acid nucleotide sequences which make them impervious to mutagenic influences, just as well as of selection for functionally necessary structures. The opposite situation in which a variety of amino acid sequences are found to occupy the same area in a set of homologous proteins, as, for example, the carboxy-terminal tetrapeptide of cytochrome c,[1-13] clearly permits the conclusion that the function of such an area, if any, is compatible with a variety of primary structures.

Differences.—The extent of variation among cytochromes c is compatible with the known phylogenetic relations of species (see Table 1). Relatively closely related species show few differences: there are only three variant residues between the horse and pig proteins. Phylogenetically distant species exhibit wider dissimilarities. The largest differences are observed between the vertebrate and yeast cytochromes c.

TABLE 1

EVOLUTION OF CYTOCHROME c

Species comparison	Number of variant residues	Divergence of lines in millions of years
Horse — Man	12	130
Horse — Pig	3	33
Horse — Chicken	12	
Pig — Chicken	10	108–150
Rabbit — Chicken	11	
Man — Chicken	14	
Horse — Tuna	19	
Pig — Tuna	17	
Rabbit — Tuna	19	184–228
Man — Tuna	21	
Chicken — Tuna	18	
Horse — Yeast	44	
Pig — Yeast	43	
Rabbit — Yeast	45	465–520
Man — Yeast	43	
Chicken — Yeast	43	
Tuna — Yeast	48	

Specific residues in certain positions appear to be characteristic of particular species or groups of closely related species. Thus, for example, a lysine in position 60, an isoleucine in position 3, a tryptophan in position 33, and an Ileu.Met sequence in positions 11 and 12, have so far been observed only in the horse, chicken, tuna, and human proteins, respectively.

Some nucleotide sequences in the deoxyribonucleic acid segment corresponding to the peptide chain of cytochrome c appear to be especially prone to mutation. Among the seven proteins, six residues (threonine, glycine, aspartic acid, glutamic acid, serine, and lysine) occur in position 89, five amino acids (glutamic acid, alanine, valine, glutamine, and asparagine) are found in position 92, while four different residues occupy each of positions 44, 58, and 60. This phenomenon might be

associated with structural or functional peculiarities of the corresponding areas in deoxyribonucleic acid. Alternatively, even with similar mutation rates for all positions, most mutations leading to changes in other locations along the peptide chain could be lethal and thus unobservable. In any case this situation is reminiscent of the genetic "hot spots" observed by Benzer[22] in bacteriophage. Clearly, the determination of the amino acid sequences of many more cytochromes *c* will be required to test the validity of the above generalizations as to invariant residues and areas, species characteristics, and highly changeable positions.

Evolution of Cytochrome c.—The qualitative aspects of the evolution of cytochrome *c* as it relates to the phylogeny of species have been discussed above. Attempts to utilize knowledge of the amino acid sequences of a set of homologous proteins in quantitative interpretations of evolution are predicated on estimates of the number of mutational events that have occurred in these proteins along two lines of evolution from the time of their divergence to date. The accuracy of such estimates is impaired by a variety of factors, among which may be cited selective evolutionary pressure, back mutation, and the lack of a necessary relation between an observed amino acid substitution and the number of steps from which it resulted.

No functional protein can possibly be a completely indifferent recorder of all chance mutational events, since some changes will lead to functionless or inefficient proteins, sooner or later to be eliminated. Nevertheless, as pointed out above, the structural requirements of a large proportion of the peptide chain of cytochrome *c* are undoubtedly compatible with a variety of sequences. Only to the extent to which this is correct will the protein act as a faithful recorder of the unit events of evolution. Moreover, the cytochromes *c* considered are functionally strictly homologous, and it is probable that selective pressure and back mutations will have effects of a similar magnitude for all of them, thus not invalidating a comparison between them.

Notwithstanding the "degeneracy"[23] of the ribonucleic acid "codes" for amino acids, an estimate of the minimal number of mutational steps that can lead from one residue to another may be possible when the nucleotide base sequences of all such "codes" will have been determined. At present one can only tentatively utilize the incomplete and possibly incorrect[24] series of base triplet sequences proposed, such as those listed by Smith,[25] Jukes,[26] or Eck.[27] Among the results which can be expected to accrue from the confrontation of a completely determined ribonucleic acid "code" with a large enough set of homologous proteins are an estimate of the amino acid sequence of the primitive protein from which other forms have developed along a particular line of evolution, and the prediction that specific residues must be present in certain positions in the proteins of extant or extinct species. Such arguments can only have a statistical validity. Thus, for example, at position 62 yeast cytochrome *c* carries an asparaginyl residue, the human, rabbit, chicken, and tuna proteins have an aspartyl residue, while the pig and horse cytochromes *c* bear a glutamyl residue. Considering it unlikely that species as divergent as tuna, chicken, rabbit, and man would all have independently changed in the identical way in the same position, it would follow that at position 62 an aspartyl residue is probably a remnant of the more primitive type of vertebrate cytochrome *c*, while the pig and the horse have mutated away from this form. Moreover, assuming for the sake of argument that the triplet "code" no. 1 proposed by Smith[25] is complete and

correct, and that only single base changes occur at each mutational step, glycine, coded by a nucleotide base sequence of UGG, is an intermediate between aspartic and glutamic acids, coded by UGA and UAG, respectively.[25] Hence, some species may carry or may have carried a glycine at this location. Within the same limitations very tentative deductions of the same general nature can be obtained for numerous positions along the peptide chain. Such deductions will become more or less probable as the correct nucleotide base sequences of amino acid "codes" are established, and the amino acid sequences of cytochromes *c* from other species are determined.

Using merely the number of variant residues to compare the seven cytochromes *c*, it is noted that the differences between them fall into groups (Table 1). Such comparisons disregard the relation of amino acid substitutions observed to the actual number of effective mutational events which occurred. Nevertheless, it appears that the number of residue differences between the cytochromes *c* of any two species is mostly conditioned by the time elapsed since the lines of evolution leading to these two species originally diverged. If this is correct, the cytochromes *c* of all mammals should be roughly equally different from the cytochromes *c* of all birds. Since fish diverged from the main stem of vertebrate evolution earlier than either birds or mammals, the cytochromes *c* of both mammals and birds should be equally different from the cytochromes *c* of fish. Similarly, all vertebrate cytochromes *c* should be equally different from the yeast protein. This is borne out by the comparison in Table 1. Clearly, the determination of the amino acid sequences of many more homologous cytochromes *c* from suitably chosen species will be required to establish the statistical validity of such a relation. It should be noted that the present results are compatible only with the commonly accepted scheme of evolution represented by series of branching lines, and are not consistent with a simultaneous formation of all species, which then proceed to accumulate mutations independently. In the latter case all the cytochromes *c* should be equally different from all others.

If elapsed time is the main variable determining the number of accumulated substitutions, it should be possible to estimate roughly the period at which two lines of evolution leading to any two species diverged. Such computations have been reported by Zuckerkandl and Pauling[28] for the various chains of human and horse hemoglobins. Using as a standard the time of separation of the lines of evolution that have led to man and horse—roughly 130 million years ago, from paleontological evidence[28]—a value of 11 million years is obtained for every residue difference along two divergent lines. Applying this value to the cytochromes *c* of other species gives the results presented in Table 1. Such times can only be expected to approximate orders of magnitude rather than exact values. An internal standard is necessary for similar estimates with every set of homologous proteins, since different genetic loci may have different rates of mutation, just as residues in different positions in a single peptide chain appear to be more or less prone to variation. In this respect cytochrome *c* as a whole would appear to be a protein which has accumulated mutations at a relatively slow rate and is therefore suitable only for the study of the evolutionary history of rather large groups of organisms. Nevertheless, the complex mutational relationships at the few highly variable locations may serve to draw distinctions between more closely related species.

The assumptions underlying such calculations are that the rate of accumulation of mutations has varied randomly[28] during evolutionary history and has not been different for homologous genetic loci in different species, even though these species exhibit a wide range of generation times. A useful test of the importance of time as the main factor in the collection of variations in cytochrome *c* would be the comparison of the amino acid sequences of the homologous proteins from species known not to have evolved morphologically for long periods with those from species which have changed rapidly, along the same general lines of evolution. Further similar tests may be possible when a valid approximation of the amino acid sequence of the primitive form of cytochrome *c* is obtained for one or more evolutionary groups of species.

Summary.—The amino acid sequences of the cytochromes *c* from horse, man, pig, rabbit, chicken, tuna, and baker's yeast show extensive identities. It is concluded that these proteins are all *homologous* structures in the evolutionary sense. Some residues and sequences in the peptide chains are invariant, while others are highly variable. Evolution exhibits a considerable degree of conservatism with regard to the remarkable clusters of basic and of hydrophobic residues in these proteins.

The cytochromes *c* from relatively closely related species show few differences, while those from phylogenetically distant species are more widely dissimilar. Particular residues in certain positions appear to be characteristic of certain species or groups of closely related species. Considering only the number of variant residues, the proteins from individual species of one class, within relatively narrow limits, are probably all equally different from those of another. The results of such comparisons are consistent with the commonly accepted over-all scheme of evolution. The extent of variation of the primary structure of the various cytochromes *c* may give rough approximations of the time elapsed since the lines of evolution leading to any two species diverged. The factors introducing errors in such estimates are discussed. Cytochrome *c* is a protein which has accumulated mutations at a relatively slow rate.

The confrontation of completely determined ribonucleic acid "codes" for amino acid residues in proteins with the amino acid sequence of a large enough set of homologous cytochromes *c* can be expected to yield an estimate of the primary structure of the primordial cytochrome *c* along one or more lines of evolution, as well as predictions of the particular residues occurring in certain positions in intermediate, extant, or extinct species.

The author is grateful to Professor K. Narita and to Dr. G. Kreil for having given him access to their studies on yeast and tuna cytochromes *c*, respectively, before publication, as well as for the generous permission to quote their results.

[1] Margoliash, E., and E. L. Smith, *Nature*, **192**, 1121 (1961).
[2] Kreil, G., and H. Tuppy, *Nature*, **192**, 1123 (1961).
[3] Margoliash, E., E. L. Smith, G. Kreil, and H. Tuppy, *Nature*, **192**, 1125 (1961).
[4] Margoliash, E., J. R. Kimmel, R. L. Hill, and W. R. Schmidt, *J. Biol. Chem.*, **237**, 2148 (1962).
[5] Margoliash, E., and E. L. Smith, *J. Biol. Chem.*, **237**, 2151 (1962).
[6] Margoliash, E., *J. Biol. Chem.*, **237**, 2161 (1962).
[7] Tuppy, H., and G. Kreil, *Monatsh. Chem.*, **92**, 780 (1962).
[8] Matsubara, H., and E. L. Smith, *J. Biol. Chem.*, **237**, PC3575 (1962).
[9] Chan, S. K., S. B. Needleman, J. W. Stewart, O. F. Walasek, and E. Margoliash, *Fed. Proc.*, **22**, 658 (1963).

[10] Kreil, G., *Z. physiol. Chem.*, in press.
[11] Narita, K., K. Titani, Y. Yaoi, H. Murakami, M. Kimura, and J. Vanecek, *Biochim. Biophys. Acta*, in press.
[12] Margoliash, E., S. B. Needleman, and J. W. Stewart, *Acta Chem. Scand.*, in press.
[13] Chan, S. K., and E. Margoliash, unpublished results.
[14] Paléus, S., and H. Tuppy, *Acta Chem. Scand.*, **13**, 641 (1959).
[15] Margoliash, E., in *Enzyme Models and Enzyme Structure*, Brookhaven Symposia in Biology No. 15 (1962), p. 266.
[16] Dus, K., R. G. Bartsch, and M. Kamen, *J. Biol. Chem.*, **237**, 3083 (1962).
[17] George, P. and R. L. J. Lyster, these PROCEEDINGS, **44**, 1013 (1958).
[18] Scouloudi, H., *Proc. Roy. Soc. (London)*, **A258**, 181 (1960).
[19] Kendrew, J. C., in *Enzyme Models and Enzyme Structure*, Brookhaven Symposia in Biology. No. 15 (1962), p. 216.
[20] Oparin, A. I., in *The Origin of Life* (New York: Academic Press, 3rd ed., 1957).
[21] Jukes, T. H., *Am. Scientist*, **51**, 227 (1963).
[22] Benzer, S., these PROCEEDINGS, **47**, 403 (1961).
[23] Crick, F. H. C., L. Barnett, S. Brenner, and R. J. Watts-Tobin, *Nature*, **192**, 1227 (1961).
[24] Crick, F. H. C., in *Progress in Nucleic Acid Research*, ed. J. N. Davidson and W. E. Cohn (New York: Academic Press, 1963), vol. 1, p. 164.
[25] Smith, E. L., these PROCEEDINGS, **48**, 677, 859 (1962).
[26] Jukes, T. H., these PROCEEDINGS, **48**, 1809 (1962).
[27] Eck, R. V., *Science*, **140**, 477 (1963).
[28] Zuckerkandl, E., and L. Pauling, in *Horizons in Biochemistry*, ed. M. Kasha and B. Pullman (New York: Academic Press, 1962), p. 198.

---

*SEPARATION OF THE TRANSFORMING AND VIRAL DEOXYRIBONUCLEIC ACIDS OF A TRANSDUCING BACTERIOPHAGE OF BACILLUS SUBTILIS**

BY S. OKUBO,† M. STODOLSKY,‡ K. BOTT,§ AND B. STRAUSS**

DEPARTMENT OF MICROBIOLOGY AND COMMITTEE ON BIOPHYSICS, THE UNIVERSITY OF CHICAGO

Communicated by George W. Beadle, August 5, 1963

In generalized transduction, a variety of genetic characters can be independently transferred from a donor to host bacteria by transducing bacteriophage. Specialized transduction is more restrictive; characteristically, only a limited segment of genetic material can be transferred by a particular bacteriophage.[1] When the *gal* locus of *Escherichia coli* is incorporated into the specialized transducing bacteriophage λ, bacterial DNA is integrated into the phage genome and replicates along with it.[2-4] The transducing particle itself is defective; singly infected *E. coli* which are transduced for the *gal* locus are not at the same time able to produce complete phages. The finding that generalized transducing phages of *E. coli* were also defective, so that transduced bacteria infected at low multiplicities were not lysogenized,[5] seemed to suggest a basic similarity between general and specialized transduction. It has been supposed[1] that transduction invariably involves recombination between bacterial and viral genetic material resulting in the insertion of bacterial DNA into the viral genome. According to this hypothesis, the difference between general and specialized transducing phages would be the number of allowed